

Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI

Year: 2020 | Citations: 335 | Authors: Sandra Wachter, B. Mittelstadt, Chris Russell

Abstract

In recent years a substantial literature has emerged concerning bias, discrimination, and fairness in AI and machine learning. Connecting this work to existing legal non-discrimination frameworks is essential to create tools and methods that are practically useful across divergent legal regimes. While much work has been undertaken from an American legal perspective, comparatively little has mapped the effects and requirements of EU law. This Article addresses this critical gap between legal, technical, and organisational notions of algorithmic fairness. Through analysis of EU non-discrimination law and jurisprudence of the European Court of Justice (ECJ) and national courts, we identify a critical incompatibility between European notions of discrimination and existing work on algorithmic and automated fairness. A clear gap exists between statistical measures of fairness as embedded in myriad fairness toolkits and governance mechanisms and the context-sensitive, often intuitive and ambiguous discrimination metrics and evidential requirements used by the ECJ; we refer to this approach as "contextual equality." This Article makes three contributions. First, we review the evidential requirements to bring a claim under EU non-discrimination law. Due to the disparate nature of algorithmic and human discrimination, the EU's current requirements are too contextual, reliant on intuition, and open to judicial interpretation to be automated. Many of the concepts fundamental to bringing a claim, such as the composition of the disadvantaged and advantaged group, the severity and type of harm suffered, and requirements for the relevance and admissibility of evidence, require normative or political choices to be made by the judiciary on a case-by-case basis. We show that automating fairness or non-discrimination in Europe may be impossible because the law, by design, does not provide a static or homogenous framework suited to testing for discrimination in AI systems. Second, we show how the legal protection offered by non-discrimination law is challenged when AI, not humans, discriminate. Humans discriminate due to negative attitudes (e.g. stereotypes, prejudice) and unintentional biases (e.g. organisational practices or internalised stereotypes) which can act as a signal to victims that discrimination has occurred. Equivalent signalling mechanisms and agency do not exist in algorithmic systems. Compared to traditional forms of discrimination, automated discrimination is more abstract and unintuitive, subtle, intangible, and difficult to detect. The increasing use of algorithms disrupts traditional legal remedies and procedures for detection, investigation, prevention, and correction of discrimination which have predominantly relied upon intuition. Consistent assessment procedures that define a common standard for statistical evidence to detect and assess *prima facie* automated discrimination are urgently needed to support judges, regulators, system controllers and developers, and claimants. Finally, we examine how existing work on fairness in machine learning lines up with procedures for assessing cases under EU non-discrimination law. A 'gold standard' for assessment of *prima facie* discrimination has been advanced by the European Court of Justice but not yet translated into standard assessment procedures for automated discrimination. We propose 'conditional demographic disparity' (CDD) as a standard baseline statistical measurement that aligns with the Court's 'gold standard'. Establishing a standard set of statistical evidence for automated discrimination cases can help ensure consistent procedures for assessment, but not judicial interpretation, of cases involving AI and automated systems. Through this proposal for procedural regularity in the identification and assessment of automated discrimination, we clarify how to build considerations of fairness into automated systems as far as possible while still respecting and enabling the contextual approach to judicial interpretation practiced under EU non-discrimination law.